doi:10.1017/S1352465819000699

MAIN

Concordance between clinician, supervisor and observer ratings of therapeutic competence in CBT and treatment as usual: does clinician competence or supervisor session observation improve agreement?

EB Caron[1,*], Michela A. Muggeo[2], Heather R. Souer[3], Jeffrey E. Pella[3] and Golda S. Ginsburg[3]

[1]Department of Psychological Science, Fitchburg State University, Fitchburg, MA 01420, USA, [2]Clarus Health Alliance, Norwich, CT 06360, USA and [3]Department of Psychiatry, University of Connecticut School of Medicine, West Hartford, CT 06119, USA

*Corresponding author. Email: ecaron3@fitchburgstate.edu

(Received 22 January 2019; revised 25 September 2019; accepted 26 September 2019)

Abstract

Background: Lowering the cost of assessing clinicians' competence could promote the scalability of evidence-based treatments such as cognitive behavioral therapy (CBT).

Aims: This study examined the concordance between clinicians', supervisors' and independent observers' session-specific ratings of clinician competence in school-based CBT and treatment as usual (TAU). It also investigated the association between clinician competence and supervisory session observation and rater agreement.

Method: Fifty-nine school-based clinicians (90% female, 73% Caucasian) were randomly assigned to implement TAU or modular CBT for youth anxiety. Clinicians rated their confidence after each therapy session ($n = 1898$), and supervisors rated clinicians' competence after each supervision session ($n = 613$). Independent observers rated clinicians' competence from audio recordings ($n = 395$).

Results: Patterns of rater discrepancies differed between the TAU and CBT groups. Correlations with independent raters were low across groups. Clinician competence and session observation were associated with higher agreement among TAU, but not CBT, supervisors and clinicians.

Conclusions: These results support the gold standard practice of obtaining independent ratings of adherence and competence in implementation contexts. Further development of measures and/or rater training methods for clinicians and supervisors is needed.

Keywords: cognitive behavioral therapy; competence; implementation; self-assessment; supervision; treatment as usual

Introduction

Mental illness impacts about one in every four children and young adults (Kazdin, 2017). A number of psychotherapies are considered effective in treating mental illness, and cognitive behavioral therapy (CBT) is among the most strongly supported. However, evidence-based treatments (EBTs) such as CBT are frequently not used in community-based practice. Further, EBTs often fail to demonstrate the same effects in community settings as in efficacy trials (e.g. Hulleman and Cordray, 2009).

One approach to improving the effects of EBTs in community settings is to enhance clinicians' treatment adherence and competence (Fixsen *et al.*, 2005) as these have been implicated as critical factors in community-based EBT failure (Hulleman and Cordray, 2009; Pellecchia *et al.*, 2015). Frequent evaluation of competence is critical to inform training efforts (Fixsen *et al.*, 2005).

Adherence and competence have traditionally been measured by independent raters who listen to complete therapy session audiotapes, making their measurement costly, time consuming and difficult to implement in community settings (Schoenwald *et al.*, 2011). One way to address these barriers may be to use clinicians' or supervisors' subjective ratings (Breitenstein *et al.*, 2010; Schoenwald *et al.*, 2011). Clinician or supervisor ratings would only be useful, however, if they were accurate (i.e. concordant with independent raters' gold standard ratings). Thus, it is critical to investigate the accuracy of clinicians' and supervisors' session-specific ratings of clinicians' competence, as well as factors that may improve their accuracy. These topics should be investigated in both EBT training and treatment as usual (TAU; in which clinicians provide their typical style of mental health care) contexts, as these settings offer complementary perspectives on bringing EBTs to the community.

Accuracy of clinician self-report

There is significant variability in the results of prior work on the accuracy of clinicians' session-specific self-assessment of competence. Agreement with independent raters, as reflected by correlations and intraclass correlation coefficients, has ranged from low to high in CBT (Brosan *et al.*, 2008; Carroll *et al.*, 1998; Loades and Myles, 2016; McManus *et al.*, 2012) and TAU (Hogue *et al.*, 2015; Hurlburt *et al.*, 2010). Generally, clinicians tend to be biased reporters who rate themselves higher in competence than independent raters (e.g. Martino *et al.*, 2009; Peavy *et al.*, 2014), a pattern that extends to community-based TAU settings (Hogue *et al.*, 2015; Hurlburt *et al.*, 2010). Among clinicians practicing CBT, however, findings have been mixed, with some work finding the typical over-estimation of competence (Brosan *et al.*, 2008; Carroll *et al.*, 1998; Rozek *et al.*, 2018) but one study finding that trainees' self-ratings were lower than independent ratings (McManus *et al.*, 2012) and another finding that clinicians were equally likely to under-estimate as over-estimate their competence (Loades and Myles, 2016).

Accuracy of supervisor report

Supervisors' reports could offer another low-cost alternative to independent ratings. That is, because supervisors already evaluate clinicians informally, asking supervisors to make formal ratings of clinicians' competence would add little additional burden, compared with independent ratings based on session recordings. The literature on supervisors' session-specific ratings of clinicians is somewhat limited, but findings have been similar to those found with clinician self-ratings. Correlations between supervisors' and independent raters' competence ratings have been variable, ranging from .14 to .61 in CBT and other EBTs (Dennhag *et al.*, 2012; Martino *et al.*, 2009; Peavy *et al.*, 2014). Additionally, with some exceptions (e.g. Rozek *et al.*, 2018), supervisors of CBT and other EBTs tend to rate clinicians' competence higher than independent raters, a pattern attributed to their loyalty to supervisees (Dennhag *et al.*, 2012; Martino *et al.*, 2009; Peavy *et al.*, 2014).

What influences accuracy of clinician and supervisor ratings?

Identifying factors that improve clinicians' and supervisors' accuracy could reduce the need for and associated costs of obtaining independent ratings. Clinicians' competence is one such factor; specifically, competent clinicians may be more self-aware regarding their skill level and therefore more accurate. In support of this idea, Brosan *et al.* (2008) found more competent CBT clinicians to be more accurate raters than less competent clinicians, and theorized that this was because competent clinicians based their self-evaluations on more relevant criteria. However, another study that compared clinician and supervisor ratings found more competent CBT clinicians to be less accurate raters (McManus *et al.*, 2012). A study among TAU clinicians found results that differed by clinic and level of clinician experience, with veteran clinicians more accurate than novice clinicians in some clinics, but the opposite true in another clinic (Hogue *et al*., 2015). Relatedly, the process of EBT training may interact with competence (Brosan *et al*., 2008), as the handful of studies in which clinicians have not shown an over-estimation bias have been conducted with trainees (Loades and Myles, 2016; Masia Warner *et al*., 2013; McManus *et al*., 2012). Clearly, more research on this topic is needed.

Use of session recordings in supervision is another factor that may promote accurate assessment of clinicians' competence. Specifically, supervisors' review of session recordings may reduce biases and improve supervisors' rating accuracy. Additionally, supervisors' session review may improve clinicians' self-rating accuracy through provision of supervisory feedback that corrects clinicians' self-perceptions (Brosan *et al*., 2008). The British Association of Behavioural and Cognitive Psychotherapies recommends using therapy session recordings in supervision and discussing discrepancies between the perceptions of clinicians and supervisors (McManus *et al*., 2012). Although this practice is thought to make supervisors' and clinicians' perceptions more accurate, this assumption should be tested.

Current study

We investigated the concordance of session-specific competence ratings made by clinicians, supervisors and independent observers in a randomized controlled trial of school-based CBT versus TAU for pediatric anxiety. Although prior work has compared different raters' concordance in CBT and other EBTs (Dennhag *et al*., 2012), to our knowledge this is the first study to examine concordance of competence ratings in CBT and TAU, offering a window into how CBT training may lead to differences in agreement about competence. In addition, although one study has examined school counselors' rating concordance with consultants (Masia Warner *et al*., 2013), this is the first study to examine school-based clinicians' concordance with independent raters.

Aims and hypotheses

Aim 1

To examine the agreement between independent observer, clinician and supervisor ratings of competence in CBT and TAU.

Hypothesis 1

Clinicians and supervisors will over-estimate competence, compared with independent observers.

Aim 2

To examine the impact of clinicians' competence on clinician and supervisor rating accuracy.

Hypothesis 2

More competent clinicians will be more accurate self-raters than less competent clinicians.

Aim 3

To examine whether supervisory session observation improves supervisors' rating accuracy.

Hypothesis 3

Session observation will be associated with higher agreement among raters.

Method

Participants

Clinicians

Participants included 24 TAU and 35 CBT clinicians. Interested clinicians were recruited from American public schools serving 5- to 18-year-olds via district supervisors, professional

Table 1. Characteristics of clinicians

Clinician	TAU ($n = 24$)	CBT ($n = 35$)	t-test or χ^2	p-value
Age: mean (*SD*)	42.6 (11.8)	43.3 (11.4)	−0.23	.82
Women	23 (96%)	30 (86%)	1.60	.21
Race				
Caucasian	17 (71%)	26 (74%)	0.05	.83
African American	4 (17%)	6 (17%)	—*	1.00
Degree				
PhD/PsyD	3 (13%)	4 (11%)	—*	1.00
Post-Masters training (CAS, CAGS, 6th year degree, EdS)	5 (21%)	8 (23%)	0.03	.87
MA/MS/MSW/LCSW/NCSP/LPC/LCPC	15 (63%)	22 (63%)	0.00	.97
Years of clinical experience: mean (*SD*)	13.7 (7.5)	15.0 (11.5)	−0.52	.61
Any participation in study supervision	16 (67%)	34 (97%)	10.23	.001
Number of supervision meetings attended: mean (*SD*)	7.4 (12.1)	11.0 (8.9)	−1.33	.19

p-value represents significance of chi-square or independent *t*-tests for differences between groups.*When expected cell counts were under 5, Fisher's exact test was used instead of chi-square.

development seminars and word of mouth. Table 1 presents information on clinicians' demographics and participation in supervision. There were no significant group differences in pre-randomization demographic characteristics, but a greater percentage of the CBT group participated in supervision, χ^2 (1) = 10.23, $p < .01$.

Supervisors

Three CBT and five TAU supervisors, recruited through word of mouth, participated. All three CBT supervisors were female doctoral-level clinical psychologists with extensive training in delivering, training and supervising CBT for pediatric anxiety. Two CBT supervisors were Caucasian and one was Biracial (Caucasian and African American). The TAU supervisors included three Caucasian females, one African American female, and one Caucasian male who were Master's level clinical social workers and clinical psychologists, selected because their therapeutic orientation was non-CBT based (e.g. psychodynamic, play therapy).

Independent raters

Eight independent raters completed ratings of therapeutic competence. All independent raters were doctoral-level psychologists in clinical psychology or a related field. Rater training consisted of co-rating and discussing several sessions with the treatment developer, and then matching within 1-point of the treatment developer on 80% of treatment adherence and competence items for two independently rated sessions.

Child participants

Session data from 203 of the 216 students enrolled in the trial (all students with a session rated by at least one rater) were used in the current study. Students were recruited through referrals from clinicians, teachers and other school staff, and parents who received a flyer. Characteristics of these students are presented in Table 2. In this sample (unlike the full sample), the TAU group included a larger percentage of African American students and the CBT group included more Caucasian students. For additional information about child participants and inclusion criteria, see Ginsburg, Pella, Pikulski, Tein and Drake (in press).

Table 2. Characteristics of child patients

Child patients	TAU ($n = 67$)	CBT ($n = 136$)	t-test or χ^2	p-value
Age: mean (*SD*)	11.5 (3.6)	10.6 (3.2)	1.59	.11
Female	33 (49%)	69 (51%)	0.04	.84
Race			9.50	.02
Caucasian	39 (58%)	93 (68%)		
African American	26 (39%)	29 (21%)		
Family income			4.80	.09
\$0–30,999	18 (27%)	20 (15%)		
\$31,000–60,999	14 (21%)	25 (18%)		
Over \$61,000	30 (45%)	77 (57%)		
Receiving free/reduced lunch	27 (40%)	35 (26%)	4.18	.04
Primary diagnosis				
Generalized anxiety disorder	36 (54%)	87 (64%)	1.97	.16
Social phobia	21 (31%)	26 (19%)	3.77	.05
Separation anxiety disorder	7 (10%)	20 (15%)	0.71	.40
Specific phobia	1 (2%)	2 (2%)	—*	1.00
Anxiety disorder not otherwise specified	2 (3%)	1 (1%)	—*	.25

p-value represents significance of chi-square or independent *t*-tests for differences between groups.*When expected cell counts were under 5, Fisher's exact test was used instead of chi-square.

Procedure

Data were collected as part of a randomized clinical effectiveness trial of a modular CBT for anxiety. Clinicians assigned to the CBT condition attended an 8-hour group training session on modular CBT, pediatric anxiety and study procedures. The majority of the training focused on the CBT modules, utilizing active learning techniques such as role plays. TAU clinicians attended a 3-hour training session on pediatric anxiety and study procedures. TAU clinicians were subsequently asked to provide the therapeutic strategies they would typically provide to anxious students; use of CBT was not explicitly restricted. Clinicians in both groups were assigned supervisors immediately after enrollment of their first study child. Supervision was optional, but available for up to 12 hours over the course of treatment. Supervision typically occurred over the telephone. Clinicians and supervisors were compensated \$35.00 and \$50.00, respectively, for every supervision hour. After each supervision session, supervisors were asked to complete a clinical supervision form which included a rating of clinicians' competence (described below).

Clinicians were instructed to provide one therapy session per week with each study child, for 12 weeks; on average, participants in the current study received 9.9 ($SD = 3.2$) sessions. Clinicians audio-recorded sessions and transferred recordings to supervisors and study staff. About 26% ($n = 395$) of available sessions ($N = 1516$) were randomly selected and coded by independent raters; 245 (62%) were from the first six sessions of therapy and 150 (38%) were from the second six sessions, reflecting the overall distribution of available sessions. After each session, clinicians were asked to complete a session summary form which included a self-rating (described below). The order of clinician self-rating and supervisor rating was not standardized. For a description of the intervention and full study procedures, see Ginsburg *et al.* (in press).

Measures

Demographic questionnaires

Clinicians completed a 15-item measure that assessed demographic and professional characteristics (e.g. age, education). Demographic information for child participants was collected through a 24-item measure completed by primary caregivers.

Supervisor ratings of competence

After each supervision session, supervisors completed an 11-point rating scale which read ‘Please rate your perception of this clinician’s competence in providing effective treatment/CBT for this anxious youth this week’ with ‘treatment’ used for TAU clinicians and ‘CBT’ used for CBT clinicians. Ratings ranged from 0 (not at all competent) to 10 (extremely competent). Supervisors also indicated whether they had listened to the audio-recorded session.

Clinician self-rated confidence

Clinicians’ self-rated confidence was measured with an item on the session summary form, completed after each therapy session. This item was designed to align with supervisor ratings of competence, but used different language and should be viewed only as a proxy for self-assessed competence, and not as a true rating of competence. The item was identical for TAU and CBT clinicians, and read, ‘How confident are you in your ability to provide effective treatment for this anxious child?’. Clinicians made ratings on an 11-point scale ranging from 0 (not at all confident) to 10 (extremely confident).

Independent rater Treatment Adherence and Therapist Competence (TATC)

The TATC (Ginsburg *et al*., 2012) included 11 items reflecting non-specific therapeutic factors (e.g. working alliance, professionalism) rated on a 4-point scale ranging from 1 (poor) to 4 (very good), which were averaged into a non-specific quality measure. Raters also completed an overall competence item aligned with the language and scale of the clinician and supervisor measures, which read, ‘Please rate your overall perception of this clinician’s competence in providing effective treatment for this anxious youth during this session’. Ratings were made on an 11-point scale ranging from 0 (not at all competent) to 10 (extremely competent). Fifteen percent of the recordings were double coded for inter-rater reliability. The one-way, random effects, single measures intraclass correlation coefficients for the competence and non-specific quality measures were .58 (95% confidence interval (CI): .39 to .73) and .63 (95% CI: .45 to .76), reflecting moderate reliability (Koo and Li, 2016). However, to facilitate comparison with inter-rater correlations (which are appropriate when measures are not identical) in the Results, the inter-rater correlations for competence and non-specific quality were r (61) = .60 and .63, both p values < .001.

Results

There were 652 TAU and 1307 CBT sessions with confidence/competence ratings from at least one rater. Figure 1 shows the distribution of ratings by raters. Although many sessions were rated only by one rater – primarily clinicians – these single rater sessions were included only in the group means and standard deviations presented in Table 3.

Aim 1: To examine the agreement between independent observer, clinician and supervisor ratings of competence in CBT and TAU

Agreement between raters was assessed in two ways. First, paired samples t-tests and Cohen’s d-scores compared average ratings between different rater groups. Second, correlations measured consistency in perceptions between rater groups. Differences in correlation magnitude were examined with Fisher r-to-z transformations (Lowry, 2001–2018). No corrections for multiple comparisons were used.

Clinicians vs independent observers

As shown in Table 3, clinicians rated their competence higher than independent observers, with a large effect for TAU clinicians (d = 1.08; t (110) = 11.40, p < .001) and a small effect for CBT clinicians (d = .30; t (259) = 4.90, p < .001).

Table 3. Means, standard deviations and correlations between competence ratings of different raters, for TAU and CBT

		Clinician-rated confidence	Supervisor-rated competence	IE-rated competence	IE: non-specific quality
	Mean (SD)	6.82 (2.11)	8.38 (1.38)	3.63 (2.31)	2.31 (0.76)
Clinician-rated confidence	7.20 (1.82)	—	.48***	.26**	.41***
Supervisor-rated competence	5.96 (2.36)	.13*	—	.21	.24
IE-rated competence	6.46 (2.05)	.19**	.17	—	.83***
IE: non-specific quality	2.98 (0.72)	.16*	.26*	.88***	—

IE, independent evaluator. Top-right corner of table (italic numbers) represents values for treatment as usual (TAU). Bottom-left corner of table (bold numbers) represents values for cognitive behavioral therapy (CBT).
*$p < .05$, **$p < .01$, ***$p < .001$.

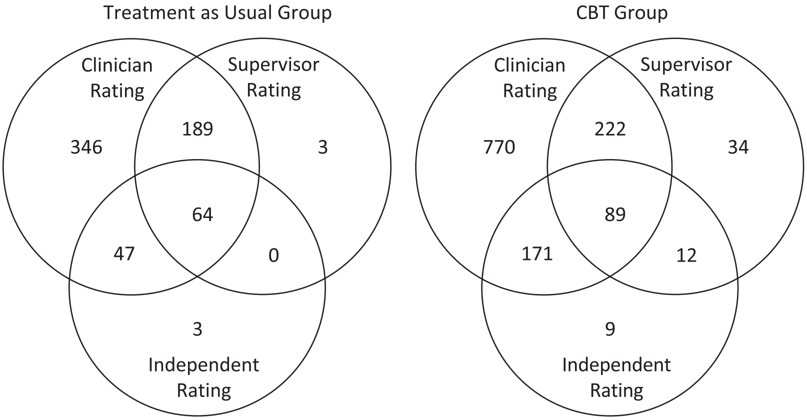

Figure 1. Distribution of rated sessions across informants.

Supervisors vs independent observers

TAU supervisors rated clinicians' competence higher than independent observers ($d = 2.32$; t (63) = 18.54, $p < .001$). In contrast, CBT supervisors rated clinicians' competence lower than independent observers ($d = -.41$; t (100) = −4.10, $p < .001$). However, while TAU supervisors consistently rated their supervisees higher than independent observers, the CBT results were driven by one CBT supervisor ('CBT Supervisor A') who rated her supervisees significantly lower than independent observers; the ratings of the other two CBT supervisors did not differ from independent observers.

Clinicians vs supervisors

CBT clinicians rated their competence higher than their supervisors ($d = .43$; t (310) = 7.51, $p < .001$). In contrast, TAU clinicians self-rated competence lower than their supervisors ($d = -.97$; t (252) = −15.43, $p < .001$).

Correlations

Inter-rater correlations are presented in Table 3 with CBT correlations in bold and TAU correlations in italics. Correlations of CBT and TAU clinicians, supervisors and independent observers were small to medium (.13 to .48). For independent raters, we included two measures of competence to examine whether the aggregated measure of non-specific quality would correlate with other raters' scores more strongly than the single-item measure of competence.

Table 4. Agreement among raters of clinicians' competence by competence level and supervisor use of session tapes

	Low competence	High competence	t-test or z-score
TAU	$n = 68$	$n = 47$	
Clinician–IE discrepancy	4.29 (*2.38*)	1.82 (*1.43*)	$t\ (107.6) = 6.79,\ p < .001$
Clinician–IE correlation	.00	.11	$z = 0.56,\ p = .58$
Supervisor–IE discrepancy	6.51 (*1.53*)	2.96 (*1.36*)	$t\ (62) = 9.24,\ p < .001$
Supervisor–IE correlation	.16	.43*	$z = 1.07,\ p = .28$
Clinician–supervisor discrepancy	2.85 (*1.64*)	1.71 (*1.04*)	$t\ (62.5) = 3.45,\ p = .001$
Clinician–supervisor correlation	.49**	.57**	$z = 0.40,\ p = .69$
CBT	$n = 134$	$n = 147$	
Clinician–IE discrepancy	2.55 (*1.70*)	1.43 (*1.38*)	$t\ (232.8) = 5.76,\ p < .001$
Clinician–IE correlation	.26**	.01	$z = -2.04,\ p = .04$
Supervisor–IE discrepancy	1.98 (*1.24*)	3.07 (*1.88*)	$t\ (94.1) = -3.50,\ p = .001$
Supervisor–IE correlation	.40**	.11	$z = -1.51,\ p = .13$
Clinician–supervisor discrepancy	2.97 (*2.11*)	2.80 (*1.98*)	$t\ (87) = 0.40,\ p = .69$
Clinician–supervisor correlation	–.12	.06	$z = 0.81,\ p = .42$
	No audio review	Audio review	t-test or z-score
TAU	$n = 20$	$n = 211$	
Supervisor–IE discrepancy	8.67 (*0.58*)	5.09 (*2.19*)	$t\ (59) = 2.81,\ p = .007$
Supervisor–IE correlation	—‡	.28*	n/a
Supervisor–clinician discrepancy	2.70 (*1.42*)	2.54 (*1.71*)	$t\ (229) = 0.42,\ p = .68$
Supervisor–clinician correlation	–.36	.54***	$z = 3.88,\ p < .001$
CBT	$n = 62$	$n = 238$	
Supervisor–IE discrepancy	1.65 (*1.46*)	2.80 (*1.71*)	$t\ (96) = -2.60,\ p = .011$
Supervisor–IE correlation	.46	.11	$z = -1.33,\ p = .18$
Supervisor–clinician discrepancy	1.68 (*1.39*)	2.60 (*1.83*)	$t\ (122.6) = -4.32,\ p < .001$
Supervisor–clinician correlation	.48***	.09	$z = -2.99,\ p = .003$

IE, independent evaluator; TAU, treatment as usual; CBT, cognitive behavioral therapy. Discrepancies are absolute values of the differences between raters; tables presents means (standard deviations, *SD*). t-tests reflect differences in discrepancy scores between groups, and z-scores reflect differences in correlations between groups. High and low competence groups were based on within-group median splits of the IE-rated single item measure of competence.

*$p < .05$, ***$p < .001$.‡Could not be computed because values for supervisor ratings were constant (i.e. all sessions were rated a 10 by supervisors).

The averaged non-specific quality measure did not correlate more strongly with clinicians' and supervisors' ratings than the independently rated single-item measure of competence (all z-values < 1.20). Interestingly, CBT Supervisor A's ratings correlated well with independent ratings, $r\ (26) = .58,\ p < .01$, whereas the other CBT supervisors had small, non-significant associations with independent observers ($r\ (18) = .06$ and $r\ (57) = .13$).

Aim 2: To examine the impact of clinicians' competence on rating accuracy

We created low and high competence groups (as rated by independent raters) with median splits within TAU (0–3 *vs* 4+) and CBT (0–6 *vs* 7+). To compare rating accuracy between low and high competence groups, we calculated discrepancy scores, which took the absolute value of the difference between the two raters' scores, representing the distance between raters' scores, irrespective of direction. Independent samples t-tests and Cohen's d-scores examined differences in discrepancy scores between high and low competence groups. Correlations assessed inter-rater consistency, and Fisher r-to-z transformations identified differences in correlation magnitude. No corrections for multiple comparisons were used.

TAU

As shown in Table 4, TAU clinicians who were rated as more competent by independent raters had lower discrepancies among raters than less competent clinicians, d values = –.84 to –2.45. Inter-rater correlations did not statistically differ between less and more competent TAU clinicians.

CBT

Findings were mixed within the CBT group. CBT clinicians who were independently rated as more competent had lower discrepancies with independent raters than less competent clinicians, $d = -.72$, but also had lower correlations with independent raters than less competent clinicians, $z = -2.04$, $p < .05$. Additionally, when CBT clinicians were more competent, their supervisors were more discrepant from independent raters than when CBT clinicians were less competent, $d = .69$. Results were not driven by any individual supervisors.

Aim 3: To examine whether supervisory session observation improves supervisors' rating accuracy

Independent samples t-tests compared discrepancy scores between sessions for which supervisors did and did not review audiotapes. Differences in inter-rater correlations were examined using Fisher r-to-z transformations. No corrections for multiple comparisons were used.

TAU

As shown in Table 4, when TAU supervisors listened to session audiotapes, they were less discrepant from independent observers than when they did not review sessions, $d = -1.66$. In addition, TAU clinicians and supervisors had stronger agreement as reflected by correlation when sessions were reviewed than when they were not reviewed, $z = 3.88$, $p < .001$. Thus, within the TAU sample, session observation appeared to improve supervisors' and clinicians' rating accuracy and concordance.

CBT

It initially appeared that when CBT supervisors listened to session audiotapes, they were more discrepant from independent observers than when they did not review sessions. However, upon examination, one supervisor ('CBT Supervisor B'), who did not observe the majority of sessions for which she provided supervision (66 of 77) and instead typically made ratings based on session discussion in supervision, skewed analyses. Specifically, CBT Supervisor B's competence ratings were less discrepant from independent observers (mean = 1.53) and marginally less discrepant from clinicians' ratings (mean = 1.42) when she did not review session audio, compared with when she did (mean = 5.00 and mean = 3.00, respectively), $t\ (16) = -3.78$, $p < .01$ and $t\ (8.59) = -2.14$, $p = .062$. In contrast, the discrepancies of CBT Supervisors A and C (who observed 258 of the 268 sessions for which they provided supervision), did not differ, whether or not sessions were observed. Thus, the findings of session review being associated with poorer agreement with independent observers do not appear generalizable. However, we did not find support for the hypothesis that session observation would improve rater agreement in the CBT sample.

Discussion

A significant barrier to EBT adoption and sustained effectiveness in community settings is costly evaluation of treatment adherence and competence by independent raters. This barrier might be lowered if clinician or supervisor ratings of clinician competence were accurate, or if factors predicting their accuracy could be identified. This study examined the concordance of competence ratings among CBT and TAU clinicians, supervisors and independent observers (Aim 1) and tested whether clinician competence and supervisor observation of sessions improved concordance (Aims 2 and 3). Although limited by the fact that measures were neither consistent between raters nor validated, the results offer an interesting signal and suggest directions for future research.

Concordance of ratings

Consistent with our hypothesis and much prior work (e.g. Dennhag *et al.*, 2012; Hogue *et al.*, 2015), in the TAU condition, both clinicians and supervisors over-estimated clinicians' competence, compared with independent observers. Although CBT clinicians also over-estimated their competence compared with independent observers, CBT supervisors instead under-estimated competence compared with independent raters. Examining inter-rater correlations, clinicians' and supervisors' agreement with independent observers was fairly low in both the CBT and TAU groups, consistent with some prior work (e.g. Dennhag *et al.*, 2012; Hurlburt *et al.*, 2010) but lower than other studies (e.g. Brosan *et al.*, 2008).

Given these findings of generally poor clinician and supervisor rating accuracy, although they are costly, independent ratings of competence are probably needed in most EBT implementation contexts. Certain supervisors may be more accurate raters than others (e.g. CBT Supervisor A), and these supervisors' competence ratings may be useful measurement tools; however, independent ratings would be necessary to identify which supervisors can make accurate ratings. Some research has found that supervisors (Reichelt *et al.*, 2003) and clinicians (Loades and Myles, 2016) can improve their rating accuracy through training, while other work has failed to find training helpful (Loades and Armstrong, 2016; Wain *et al.*, 2015). Future work should explore individual differences in rating accuracy and ways to improve it.

Another direction for future research is to identify aspects of measurement that impact clinicians' and supervisors' rating accuracy, to promote the development of adherence and competence measures on which clinicians and supervisors can be reliable with independent observers. For example, clinicians can be good reporters of clearly defined elements of adherence (e.g. coverage of a specific CBT module, assignment of homework) based on their informal, verbal report to supervisors during consultation (Ward *et al.*, 2013). Additionally, measures of adherence and competence based on tallying instances of clinicians' specific verbal behaviors appear to promote clinicians' and supervisors' inter-rater reliability to a greater extent than the broad competence ratings used in this study (Caron and Dozier, 2019). Although we failed to find that the 11-item non-specific quality scale was more highly correlated with clinicians' and supervisors' ratings than the single-item competence rating, future research should examine the benefits of multi-item scales for various raters. For example, the multi-item scale may have calibrated independent raters' use of the single-item measure, as they typically completed the non-specific ratings prior to scoring overall competence, and correlations between the measures were high (r values > .80).

Impact of competence on rating accuracy

Consistent with prior work (Brosan *et al.*, 2008; McManus *et al.*, 2012), findings regarding the impact of clinicians' competence on rating accuracy were mixed. Specifically, as hypothesized, more competent TAU clinicians had lower discrepancies with supervisors and independent raters than less competent clinicians. These results were consistent with the results of Brosan *et al.* (2008), who recruited a sample of CBT clinicians practicing independently; in both of these groups that were practicing as usual, more competent clinicians were more accurate in assessing their competence. In the CBT group, however, findings were mixed, with some results suggesting that more competent CBT clinicians were less accurate self-raters and also received less accurate supervisor ratings. In line with these findings, McManus *et al.* (2012) found that more competent CBT trainees tended to be less accurate self-raters; similar to the sample of McManus *et al.* (2012), the CBT clinicians in this study were learning to implement a therapy new to them. In the context of learning and practicing new material, perhaps more competent therapists are more self-critical and therefore less accurate in assessing their competence than less competent clinicians. In sum, the current results suggest that more competent clinicians who are practicing as usual and are not learning new practices are likely to be more accurate self-raters than others, whereas competent clinicians learning EBTs may struggle to accurately evaluate their performance. However, more research is needed to further explore training stage as a potential moderator of clinician and supervisor rating accuracy, as well as other possible reasons for discrepant findings, including methodological and rater differences between studies (McManus *et al*., 2012).

Impact of supervisory session observation on rating accuracy

Consistent with hypotheses, supervisors' perceptions of TAU clinicians' competence were more in line with independent observers when supervisors reviewed session audiotapes. These results suggest that listening to session audiotapes improved the accuracy of TAU supervisors' perceptions of clinicians, probably because supervisors and independent raters were basing ratings on the same material (i.e. what occurred during the session) as opposed to clinicians' self-report to supervisors about sessions. In addition, the correlation between TAU supervisors' and clinicians' ratings of competence was higher when sessions were reviewed than not. These results suggest that session review helped to align clinicians' and supervisors' perceptions of performance, perhaps because it allowed supervisors to provide feedback that corrected clinicians' self-perceptions (Brosan *et al*., 2008). Thus, it appears that session observation in supervision, although under-utilized in community settings, may be helpful, consistent with recommendations in the field (e.g., Hurlburt *et al*., 2010).

We cannot draw conclusions about the effects of session observation in the CBT sample, as large supervisor differences influenced analyses, and the small supervisor sample size limited generalizability. Supervision in the current study was fairly unstructured (e.g. session audio review was optional), which probably contributed to these large CBT supervisor differences. Our results suggest that individual level effects of CBT supervisors should be considered in future research, and CBT supervisors' evaluations of clinicians' competence may be affected when deviating from their typical supervisory practices.

Strengths and limitations

The current study had several limitations related to measurement. First, the only measure shared across both TAU and CBT groups, and all three sets of raters, was a single-item measure of therapeutic competence/confidence. These measures were not validated, which would have allowed comparison of this study with other samples. Although the items for supervisors and independent raters had very similar wording and directly assessed 'competence,' the item for clinicians asked about 'confidence.' While we thought that indirectly assessing competence with this language would be more comfortable for clinicians, it is unknown how assessing confidence instead of competence affected our results, and it is possible that directly assessing competence would have led to higher agreement with independent raters. Future work should include validated multi-item measures of both adherence and competence that are identical across raters. Another limitation of the measurement was that the order of ratings was not standardized; specifically, clinicians could have completed their ratings before or after supervision, and could have been independent from or influenced by supervisors' input. Additionally, although sessions were selected randomly for independent rating, clinician and supervisor ratings depended on their completion of measures, and were not available when clinicians did not participate in supervision or complete session summary forms, which may have biased results. Relatedly, correlations between raters were conducted on different subsamples of tapes. Furthermore, supervisors' ratings were probably biased by prior knowledge and global perceptions of clinicians. Independent ratings may also have been influenced in these ways; for example, one author who rated many tapes had a 'favorite' TAU clinician. In addition, inter-rater reliability on the TATC among trained raters was only moderate, setting a low ceiling for what we could expect from clinicians and supervisors. A final limitation is that independent raters probably believed in the efficacy of CBT and influenced the size of the TAU group's discrepancies from independent raters. Although we chose not to directly compare the CBT and TAU groups because of this issue, future work may benefit from using unbiased raters and measures.

The study also had several strengths. First, the session-level sample size was large, and allowed analyses that split the sample into CBT and TAU groups, and further splitting of the sample by competence, audio review and CBT supervisor. Additionally, the randomization of clinicians to conditions allowed us to infer that different patterns of results between the groups, such as the mixed findings with regard to impact of clinician competence on rating accuracy, are due to training condition. Although use of CBT was not restricted in the TAU condition, it was low; specifically, 14% of the independently rated TAU sessions were identified as utilizing cognitive behavioral strategies, demonstrating group differentiation (Ginsburg *et al.*, 2019). In addition, many randomized trials do not include supervision in a TAU condition, but because supervision was provided, TAU supervisors' ratings were available, and the effect of audio review of sessions by both CBT and TAU supervisors could be examined, a novel contribution to the literature. Finally, the study clinicians were recruited from schools and intervened with children in the school setting. Understanding strategies to support the process of training community-based clinicians in EBTs is a critical target for the field, and schools represent a particularly important context for community-based implementation of EBTs (Farmer *et al.*, 2003).

Conclusions

In summary, this study found that both CBT and TAU clinicians and supervisors had fairly low correlations with independent observers. Hypotheses tended to be supported in the TAU group, but not the CBT group, which may reflect a destabilizing impact associated with learning a new EBT. Specifically, higher levels of clinician competence and supervisory session observation appeared to promote alignment between the perspectives of TAU clinicians, supervisors and independent raters. However, findings were variable among CBT clinicians and supervisors. More work is needed to identify measures of adherence and competence on which clinicians and supervisors can achieve acceptable reliability with observers, and training or feedback procedures to promote clinicians' and supervisors' rating accuracy. Until these measures and procedures are validated, independent observer ratings should continue to be used in implementation contexts as the demand for EBTs increases.

Acknowledgements. We are grateful for the hard work and support of many research assistants and school personnel throughout the course of the study. We would also like to thank all the clinicians, students and parents who participated in the study. We are especially grateful to Kelly Drake, Katie Herzig, Courtney Keeton, Angela Allen Peck, Beverly Coker, Sean Cronin, Ann Oidick, Nina Rosenzwog, Jennifer Nail, Sarah Williams, Amy Hale, Elizabeth Casline, Paige Pikulski, Jamie LoCurto, Brian Padilla, and many other students and fellows for their assistance with implementing this study.

Financial support. This study was supported by the Institute of Educational Sciences, U.S. Department of Education (grant number R324A120405) awarded to Dr. Golda S. Ginsburg. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Conflicts of interest. EB Caron, Michela Muggeo, Heather Souer, Jeffrey Pella and Golda Ginsburg have no conflicts of interest with respect to this publication.

Ethical statements. The authors have abided by the Ethical Principles of Psychologists and Code of Conduct as set out by the American Psychological Association. Ethical approval of procedures was provided by the University of Connecticut School of Medicine Institutional Review Board (submission no. 012737).

References

Breitenstein, S. M., Gross, D., Garvey, C. A., Hill, C., Fogg, L., & Resnick, B. (2010). Implementation fidelity in community-based interventions. *Research in Nursing & Health*, *33*, 164–173. doi: 10.1002/nur.20373

Brosan, L., Reynolds, S., & Moore, R. G. (2008). Self-evaluation of cognitive therapy performance: do therapists know how competent they are? *Behavioural and Cognitive Psychotherapy*, *36*, 581–587. doi: 10.1017/S1352465808004438

Caron, E., & Dozier, M. (2019). Effects of fidelity-focused consultation on clinicians' implementation: an exploratory multiple baseline design. *Administration and Policy in Mental Health and Mental Health Services Research*, *46*, 445–457. doi: 10.1007/s5488-019-00924-3

Carroll, K., Nich, C., & Rounsaville, B. (1998). Utility of therapist session checklists to monitor delivery of coping skills treatment for cocaine abusers. *Psychotherapy Research*, *8*, 307–320. doi: 10.1080/10503309812331332407

Dennhag, I., Gibbons, M. B. C., Barber, J. P., Gallop, R., & Crits-Christoph, P. (2012). Do supervisors and independent judges agree on evaluations of therapist adherence and competence in the treatment of cocaine dependence? *Psychotherapy Research*, *22*, 720–730. doi: 10.1080/10503307.2012.716528

Farmer, E. M., Burns, B. J., Phillips, S. D., Angold, A., & Costello, E. J. (2003). Pathways into and through mental health services for children and adolescents. *Psychiatric Services*, *54*, 60–66. doi: 10.1176/appi.ps.54.1.60

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation research: a synthesis of the literature (FMHI Publication No. 231)*. Tampa: University of South Florida, Louis de la Parte Florida Mental Health Institute, The National Implementation Research Network. Retrieved from: https://fpg.unc.edu/node/4445

Ginsburg, G. S., Becker, K. D., Drazdowski, T. K., & Tein, J. Y. (2012). Treating anxiety disorders in inner city schools: results from a pilot randomized controlled trial comparing CBT and usual care. *Child & Youth Care Forum*, *42*, 1–19. doi: 10.1007/s5566-011-9156-4

Ginsburg, G. S., Muggeo, M., Caron, E., Souer, H. R., & Pikulski, P. J. (2019). Exploring treatment as usual for pediatric anxiety disorders among school-based clinicians. *School Mental Health*, *11*, 719–727.

Ginsburg, G. S., Pella, J. E., Pikulski, P. J., Tein, J. T., & Drake, K. L. (in press). School-based treatment for anxiety research study (STARS): a randomized controlled effectiveness trial. *Journal of Abnormal Child Psychology*.

Hogue, A., Dauber, S., Lichvar, E., Bobek, M., & Henderson, C. E. (2015). Validity of therapist self-report ratings of fidelity to evidence-based practices for adolescent behavior problems: correspondence between therapists and observers. *Administration and Policy in Mental Health and Mental Health Services Research*, *42*, 229–243. doi: 10.1007/s5488-014-0548-2

Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: the role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness*, *2*, 88–110. doi: 10.1080/19345740802539325

Hurlburt, M. S., Garland, A. F., Nguyen, K., & Brookman-Frazee, L. (2010). Child and family therapy process: concordance of therapist and observational perspectives. *Administration and Policy in Mental Health and Mental Health Services Research*, *37*, 230–244. doi: 10.1007/s5488-009-0251-x

Kazdin, A. E. (2017). Addressing the treatment gap: a key challenge for extending evidence-based psychosocial interventions. *Behaviour Research and Therapy*, *88*, 7–18. doi: 10.1016/j.brat.2016.06.004

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*, 155–163. doi: 10.1016/j.jcm.2016.02.012

Loades, M. E., & Armstrong, P. (2016). The challenge of training supervisors to use direct assessments of clinical competence in CBT consistently: a systematic review and exploratory training study. *The Cognitive Behaviour Therapist*, *9*. doi: 10.1017/S1754470X15000288

Loades, M. E., & Myles, P. J. (2016). Does a therapist's reflective ability predict the accuracy of their self-evaluation of competence in cognitive behavioural therapy? *The Cognitive Behaviour Therapist*, *9*. doi: 10.1017/S1754470X16000027

Lowry, R. (2001–2018). Significance of the difference between two correlation coefficients. *VassarStats: Website for Statistical Computation*. Retrieved from: vassarstats.net/index.html

Martino, S., Ball, S., Nich, C., Frankforter, T. L., & Carroll, K. M. (2009). Correspondence of motivational enhancement treatment integrity ratings among therapists, supervisors, and observers. *Psychotherapy Research*, *19*, 181–193. doi: 10.1080/10503300802688460

Masia, C. W., Brice, C., Esseling, P. G., Stewart, C. E., Mufson, L., & Herzig, K. (2013). Consultants' perceptions of school counselors' ability to implement an empirically-based intervention for adolescent social anxiety disorder. *Administration and Policy in Mental Health*, *40*, 541–554.

McManus, F., Rakovshik, S., Kennerley, H., Fennell, M., & Westbrook, D. (2012). An investigation of the accuracy of therapists' self-assessment of cognitive-behaviour therapy skills. *British Journal of Clinical Psychology*, *51*, 292–306. doi: 10.1111/j.2044-8260.2011.02028.x

Peavy, K. M., Guydish, J., Manuel, J. K., Campbell, B. K., Lisha, N., Le, T., . . . & Garrett, S. (2014). Treatment adherence and competency ratings among therapists, supervisors, study-related raters and external raters in a clinical trial of a 12-step facilitation for stimulant users. *Journal of Substance Abuse Treatment*, *47*, 222–228. doi: 10.1016/j.jsat.2014.05.008

Pellecchia, M., Connell, J. E., Beidas, R. S., Xie, M., Marcus, S. C., & Mandell, D. S. (2015). Dismantling the active ingredients of an intervention for children with autism. *Journal of Autism and Developmental Disorders*, *45*, 2917–2927. doi: 10.1007/s5803-015-2455-0.

Reichelt, F. K., James, I. A., & Blackburn, I. M. (2003). Impact of training on rating competence in cognitive therapy. *Journal of Behavior Therapy and Experimental Psychiatry*, *34*, 87–99. doi: 10.1016/S0005-7916(03)00022-3

Rozek, D. C., Serrano, J. L., Marriott, B. R., Scott, K. S., Hickman, L. B., Brothers, B. M., … **& Simons, A. D.** (2018). Cognitive behavioural therapy competency: pilot data from a comparison of multiple perspectives. *Behavioural and Cognitive Psychotherapy*, *46*, 244–250. doi: 10.1017/S1352465817000662

Schoenwald, S. K., Garland, A. F., Chapman, J. E., Frazier, S. L., Sheidow, A. J., & Southam-Gerow, M. A. (2011). Toward the effective and efficient measurement of implementation fidelity. *Administration and Policy in Mental Health and Mental Health Services Research*, *38*, 32–43. doi: 10.1007/s5488-010-0321-0

Wain, R. M., Kutner, B. A., Smith, J. L., Carpenter, K. M., Hu, M. C., Amrhein, P. C., & Nunes, E. V. (2015). Self-report after randomly assigned supervision does not predict ability to practice Motivational Interviewing. *Journal of Substance Abuse Treatment*, *57*, 96–101. doi: 10.1016/j.jsat.2015.04.006

Ward, A. M., Regan, J., Chorpita, B. F., Starace, N., Rodriguez, A., Okamura, K., … & **Research Network On Youth Mental Health** (2013). Tracking evidence based practice with youth: validity of the MATCH and Standard Manual Consultation Records. *Journal of Clinical Child & Adolescent Psychology*, *42*, 44–55. doi: 10.1080/15374416.2012.700505

Cite this article: Caron E, Muggeo MA, Souer HR, Pella JE, and Ginsburg GS. Concordance between clinician, supervisor and observer ratings of therapeutic competence in CBT and treatment as usual: does clinician competence or supervisor session observation improve agreement? *Behavioural and Cognitive Psychotherapy*. https://doi.org/10.1017/S1352465819000699